Secrets of RLHF in Large Language Models Part II: Reward Modeling

Year: 2024 | *Citations: 135* | *Authors: Bing Wang, Rui Zheng, Luyao Chen*

Abstract

Reinforcement Learning from Human Feedback (RLHF) has become a crucial technology for aligning language models with human values and intentions, enabling models to produce more helpful and harmless responses. Reward models are trained as proxies for human preferences to drive reinforcement learning optimization. While reward models are often considered central to achieving high performance, they face the following challenges in practical applications: (1) Incorrect and ambiguous preference pairs in the dataset may hinder the reward model from accurately capturing human intent. (2) Reward models trained on data from a specific distribution often struggle to generalize to examples outside that distribution and are not suitable for iterative RLHF training. In this report, we attempt to address these two issues. (1) From a data perspective, we propose a method to measure the strength of preferences within the data, based on a voting mechanism of multiple reward models. Experimental results confirm that data with varying preference strengths have different impacts on reward model performance. We introduce a series of novel methods to mitigate the influence of incorrect and ambiguous preferences in the dataset and fully leverage high-quality preference data. (2) From an algorithmic standpoint, we introduce contrastive learning to enhance the ability of reward models to distinguish between chosen and rejected responses, thereby improving model generalization. Furthermore, we employ meta-learning to enable the reward model to maintain the ability to differentiate subtle differences in out-of-distribution samples, and this approach can be utilized for iterative RLHF optimization.